IJSPT

ORIGINAL RESEARCH

THE FUNCTIONAL MOVEMENT SCREENING (FMS)™: AN INTER-RATER RELIABILITY STUDY BETWEEN RATERS OF VARIED EXPERIENCE

Heather Gulgin, PhD, ATC[1]
Barbara Hoogenboom, PT, EdD, SCS, ATC[1]

ABSTRACT

Background: Previous researchers have reported on the reliability of the scoring of the FMS™ movement screens. Those authors have reported good to excellent inter-rater reliability between paired raters of similar experience level (either novice or expert), but no comparisons of inter-rater reliability exist between a novice and an expert.

Purpose: The purpose of this investigation was to examine the inter-rater reliability of the scoring of the FMS™ between trained novices and an expert rater using video records.

Methods: Twenty healthy college students participated. Each participant performed the series of seven functional movement screens. Four raters (three novices and one expert) independently scored the seven FMS™ tests by watching video recordings of the movements..

Results: The mean total FMS™ score for all subjects was 14.6 $\pm$ 1.9, and was not significantly different between raters ($p = 0.136$). For the individual tests, half of them had perfect agreement, while the other half ranged from slight to moderate agreement (33-66%).

Conclusion: Total FMS™ scores were similar among the raters, and the inter-rater reliability for a majority of the individual tests had as strong agreement despite the various level of experience of the raters scoring the FMS™ tests.

Clinical Relevance: Although there was mostly moderate to perfect agreement among raters, the level of experience of the rater scoring the FMS™ should be considered, as it appears that the expert rater was more critical than novice raters in the interpretation of the scoring criteria.

Key Words: Functional movement screen, reliability

Level of Evidence: Level 3

CORRESPONDING AUTHOR

Heather Gulgin, PhD, ATC
Movement Science Department
Grand Valley State University
1 Campus Drive, MAK B2-222
Allendale, MI 49401
Email: gulginh@gvsu.edu

[1] Grand Valley State University, Grand Rapids, MI, USA

INTRODUCTION

Multiple factors including previous injury, conditioning, and skill level may predispose athletes to injury. Compensatory movements may be related to muscle tightness, muscular weakness, and altered or poor neuromuscular patterns.[1-3] For decades rehabilitation professionals have used muscle length tests, manual muscle tests, and range of motion assessments to attempt to determine not only the status of their athletes with regard to impairments and dysfunction, but also regarding the readiness of injured athletes to return to participation. Such an approach is likely to fall short of assessing the function of the entire kinetic chain. In order to attempt to screen for factors relating to injury, professionals have begun to assess function using foundational movement patterns that require coordinated utility of multiple joints and their movements.[1-3] A series of movement patterns, when used as a test battery, may help determine motor dysfunction if it exists.

The term "regional interdependence" is used to describe the relationship between regions of the body and how dysfunction in one region may contribute to dysfunction in another region.[4,5] In fact, it is becoming apparent that what may appear to be an isolated injury or dysfunction may have far reaching effects in regions away from the injury site.[6-11] Nadler et al[10] demonstrated that rehabilitation after injury should not be isolated to the injured region, rather, it should address the athlete as a whole in order to return the athlete to the highest level of function.[12] Thus, it is important that a regional approach to movement, function, and rehabilitation be employed by those treat patients of all abilities.

A system that can be used to assess fundamental functional performance, The Functional Movement Screen (FMS™) has been described by Cook and colleagues.[2-3] If valid and reliable, tests that assess multiple domains of function may enhance the ability of professionals to identify athletes at risk for injury.[13] The FMS™ is comprised of seven multi-joint movement patterns designed to assess and rate functional movement by incorporating upper and lower extremity sequenced movements that require concurrent coordination of both stability and movement of the trunk and pelvis. These tests incorporate many facets of human movement including strength, motor control, balance, and symmetry. Cook et al[1-3] have suggested that fundamental movements, such as those that comprise the FMS™, are basic movement skills that may be related to a wide variety of movement patterns used in ADL's and sports. The seven FMS™ tests and the criteria for scoring their performance have been well described in previous studies.[2,3,13-15]

Preliminary investigations by Kiesel et al[16] and Chorba et al[14] described the use of the FMS™ for screening athletes and attempted to determine the predictive value of the FMS™ related to injury. Kiesel et al[16] determined that athletes who scored 14 or less on the FMS™ possessed dysfunctional movement patterns that may correlate with greater risk of injury. Chorba et al[14] examined female collegiate athletes and found that those who scored less than 14 on the FMS™ had an approximate four-fold increase in risk (odds ratio 3.85-4.58) of lower extremity injury throughout the course of a season. There was a significant correlation between low-scoring athletes and injury ($p = 0.021$, $r = 0.76$) suggesting that the FMS™ may be able to successfully predict which female athletes, without a history of previous musculoskeletal injury, would be injured over the course of a season.[14]

Minick et al[13] described the inter-rater reliability of the FMS™ between pairs of novice and expert raters, all trained in scoring of the FMS™. Video recordings of the FMS™ tests were given to the raters to score. Statistical analysis using the weighted Kappa Statistic suggested excellent or substantial agreement between raters on the majority of the tests. The authors suggested that the FMS™ could be reliably used to assess movement patterns of athletes and recognize which individuals may be at risk for injury.

Onate et al[17] and Smith et al[18] assessed the inter-rater reliability of the FMS™ among raters of different levels of training and experience using real-time analysis. They found excellent (ICC = 0.98) to good (ICC's 0.87-0.89) inter-rater reliability for total scores, respectively. Inter-rater reliability of individual FMS™ movements was also reported by Smith et al, and ranged from fair to good (ICC's 0.30-0.89), however, percent agreement between raters was not reported.

Gribble et al[19] and Smith et al[18] and assessed the intra-rater reliability of total FMS™ scores using both video-taped (Gribble) and real-time (Smith) assess-

ments. They reported reliability between raters of different training and experience as ranging from ICC = 0.37-0.95[19] to ICC = 0.81-0.91.[18] In the study by Gribble et al[19] athletic trainers with an average of 6.41 years of experience had the strongest intra-rater reliability (ICC = 0.95) while the non-certified student athletic trainers had only poor intra-rater reliability (ICC = 0.37). Onate et al[17] also reported on intra-rater reliability of both FMS™ among certified and non-certified raters with widely variant levels of experience. Their findings demonstrated poor to good reliability of the individual FMS™ movements with Kappa values ranging from 0.16-0.84.

Schneiders et al[20] published a description of normative FMS™ scores using active male and female subjects. Their mean composite FMS™ score was 15.7 out of 21 (CI 15.4-15.9) with no significant difference between females and males. Additionally, they used a sub-group of 28% of their sample, to determine reliability of scoring between the two trained and experienced raters who participated. However, the authors did not clarify what comprised trained or experienced in their raters. The authors were able to demonstrate excellent inter-rater reliability for the total FMS™ score (ICC = 0.971), and Kappa scores ranging from 0.73-1.00 for the individual FMS™ tests, demonstrating substantial to excellent inter-rater reliability.

In summary, Schneiders et al[17] found excellent inter-rater reliability among two raters with the same level of experience, and Minick et al[13] found excellent inter-rater reliability when comparing pairs of novice raters or pairs of expert raters (also same levels of experience). Two groups of authors[18,19] have reported a variety of intra-rater reliability values, and only Smith et al[18] reported inter-rater reliability of scoring between raters of different levels of training and experience using real time analysis. Thus, whether raters of various experience levels demonstrate equally high inter-rater reliability remains incompletely explored. The purpose of this investigation was to examine the inter-rater reliability of the scoring of the FMS™ between trained novices and an expert rater using video records.

METHODS

Subjects

Twenty college-aged subjects (Table 1) were recruited by word of mouth. Subjects were active student volunteers who were free of lower extremity injury or dysfunction, and able to participate in all normal activities at the time of the study. Prior to participation, subject's signed an informed consent form approved by the Human Subjects Review Board.

Table 1. *Subject Demographics.*

Subject	Height (cm)	Weight (kg)	Age (yrs)
Males (n = 10)	179.04 ± 7.03	73.96 ± 10.09	20.44 ± 2.12
Females (n = 10)	170.28 ± 4.94	60.66 ± 7.88	19.62 ± 0.91

Procedures

Subjects reported to the Biomechanics & Human Performance Lab on one occasion for the testing protocol. The subject's height and weight were measured. Prior to the subject performing a given functional movement screen, the investigator first demonstrated the movement, while providing standard instructional cues developed from the descriptions provided by Cook et al.[2,3] on how to generally perform the movement. Each subject performed the seven functional movement tests, in the order described by Cook et al.[2,3] while being videotaped from sagittal and frontal views (Sony Handycam, Model DCR-TRV70, Sony Corporation, Tokyo, Japan). When appropriate, the functional movement screens were performed separately on the right and left sides.

Three novice raters (third-year physical therapy students, all certified in FMS™ scoring) and one expert rater (Formal training before certification existed, 3 years regular experience with FMS™ scoring), independently assessed the performance of the FMS™ tests performed by the subjects, by watching video recordings of the movements at normal speed. No slow motion viewing was allowed to attempt to replicate real-time scoring. The raters were allowed to replay the video. They used the FMS™ standard scoring criteria originally developed and published for the seven screening tests.[1-3] All scores were recorded on a data collection sheet, and when bilateral scores differed the lower of the two was used for the total score, as per the test instructions. A total score of 21 is the highest score a participant can achieve.

STATISTICAL METHODS

The Fisher's Exact test was used to compare scores among raters, since the data was non-parametric,

and not all assumptions for the Chi-squared test were met. This test allows for each novice rater to be compared to the expert rater, or look for agreement among a pair of raters. The statistical test detected whether significant discrepancies existed in the scoring for a given FMS™ test between raters. The ICC could not be used for the data analysis of the individual movements because the interval between scores awarded on the FMS™ test cannot be considered equal. The use of a weighted Kappa statistic as described by Minick et al[13] is reserved for examining agreement between paired raters. Since the study design called for more than one novice rater to be compared to the expert rater the authors could not compare to the expert rater by using a weighted Kappa. Mean total FMS™ scores were examined using a one-way ANOVA to determine whether significant differences existed between raters ($p < 0.05$). Mean total scores were also examined using the Intraclass correlation coefficient ($ICC_{3,1}$) in order to determine inter-rater reliability of these scores between raters.

RESULTS

The mean total score on the FMS™ for all subjects by all raters was 14.64 (SD 1.9) and the mean total FMS™ score for each rater is shown in Table 2. Although rater four (the expert) had a slightly lower total mean FMS™ score than the other three raters, there were no significant differences in mean scores between raters ($p = 0.14$). The ICC for mean total scores was 0.88 (95% CI .767-.948) indicating good to excellent overall consistency between raters (Table 2).

Of the twelve separate tests that comprise the FMS™ six tests demonstrated perfect agreement among the raters (Table 3). Three of the twelve tests (Hurdle step L, Hurdle step R, Shoulder mobility L) demonstrated moderate agreement (66%), and three tests (Squat, Rotary Stability R, Rotary Stability L) demonstrated slight agreement (33%).

Table 2. *Composite FMS™ scores and Inter-rater reliability for all subjects (n = 20)*

Rater	Composite FMS™ Score Mean ± SD	95% CI	Range (min-max)	ICC
Rater 1	15.25 ± 1.71	14.5-16.0	11-18	
Rater 2	14.40 ± 1.66	13.6-15.1	11-17	
Rater 3	14.95 ± 2.03	14.0-15.9	11-19	
Rater 4	13.95 ± 2.03	13.0-14.9	11-19	
Total	14.64 ± 1.90	14.2-15.0	10-19	.882
Non-significant difference (p = 0.14)				

Table 3. *Inter-rater Agreement (Fisher's Exact) between Novice Raters & Expert Rater for Individual Tests Comprising the FMS™*

Test	Novice 1 vs. Expert	Novice 2 vs. Expert	Novice 3 vs. Expert	% Agreement
Squat	.007*	.299	<.001*	33%
Hurdle R	.092	.458	.285	100%
Hurdle L	.028*	.221	.115	66%
Lunge R	.040*	.333	.751	66%
Lunge L	.333	.095	.190	100%
Shoulder Mobility R	.698	.406	.519	100%
Shoulder Mobility L	.633	.054	<.001*	66%
ASLR R	.621	.097	.285	100%
ASLR L	.374	.129	.384	100%
Trunk Stability	1.00	.482	.756	100%
Rotary R	<.001*	.091	<.001*	33%
Rotary L	.021*	.405	.015*	33%

ASLR = active straight leg raise
R = right
L = left
*Fisher's Exact significant at p = 0.05

In order to examine the tests with slight agreement for inter-rater reliability, further analysis of scoring was achieved by examining the actual score distribution among the novices and expert to determine if experience level may have allowed for more critical or lenient interpretation of scoring of FMS™ tests. For each of the tests that only had 33% agreement, the expert rater scored the performance lower than the novices.

DISCUSSION

Total FMS™ Scores

Total FMS™ scores generated by the four raters were not significantly different, demonstrating the ability of four different raters (one expert, three novices) to score the group of tests similarly. Likewise the scores of the four raters demonstrated good to excellent correlation ($ICC_{3,1} = 0.882$). This is consistent with the results of Schneiders et al[20] whose raters total FMS™ scores were closely related (ICC = 0.971), the inter-rater reliability reported by Smith et al[18] (ICC's ranging between 0.87-0.89), and inter-rater reliability demonstrated by the raters in the Onate et al[17] study (ICC's 0.92-0.98). Schneiders et al[20] did not report the level of clinical experience of their raters, only that they were experienced, and the

raters in the Onate et al.[17] study varied between an ATC with four years of experience, and other moderately experienced raters, as well as a true novice who had only read the FMS™ manual. The raters in the current study utilized video analysis while raters in the other three studies used real-time analysis, thus accounting for the greater scrutiny of the raters in the current study, and slightly lower inter-rater reliability. Because the rater's total FMS™ scores in the current study were very similar, and similar to others reported in the literature, it would have been helpful to examine total FMS™ found in the Minick et al[13] study. Unfortunately, although Minick et al[13] reported excellent agreement among raters, they did not report total FMS™ scores to allow comparisons between total scores between raters in their study and those who completed the ratings in the current study.

The mean total FMS™ score of the current study group of healthy male and female subjects (14.64, SD 1.90) was similar to the male and female athletes in the Smith study (14.3, SD 1.5) and the female athletes in the Chorba et al[14] study (14.3, SD 1.77). However mean total FMS™ scores for the subjects in the current study were lower than those reported by Schneiders et al[20] (15.7; 95% CI 15.4-15.9), and Onate et al[17] during their inter-rater reliability analysis (16.75; 95% CI 15.81-17.74).

Inter-rater Agreement

Data in the current study indicate that half of the FMS™ tests demonstrated perfect agreement in scoring (Table 3) when comparing between novice and expert raters. The purpose of the current study was to examine the reliability of scoring between FMS™ raters of varied experience levels, as compared to earlier studies that used homogenous pairs of raters or raters with widely variant experience. The current results support most of the reliability (agreement) results presented by Minick et al, although the some of the individual tests showed less agreement among raters than those demonstrated by the raters in Minick et al.[13] Despite the similar methodologies used in the two studies, the current study compared scoring between novice and expert raters (3 pairs; 1 novice vs. 1 expert in each pair), whereas Minick et al[13] examined agreement between a single pair of novice raters and a single pair of expert raters. It is possible that the ability of the novice raters were more similar to each other in scoring the FMS™ tests, and likewise greater similarities existed in scoring between expert raters, thus allowing for a higher inter-rater agreement by Minick et al.[13]

Three individual tests demonstrated moderate agreement between raters, while the remaining three individual tests demonstrated slight agreement. Upon further examination, in the tests that demonstrated poorer agreement, the authors found that the expert was more critical (scored participants lower) than the novices the majority of the time. It is possible that the expert rater became more critical with experience. Thus, in the current study, the authors see a slight difference of scoring based on the experience level of rater, whereas the previous studies could not compare raters with a heterogenous level of experience. This could explain the overall poorer levels of agreement in the current study as compared to others.

The individual FMS™ tests with the lowest inter-rater reliability in the current study, the Squat and Rotary Stability tests, also had the lowest level of agreement between raters in both the Minick et al[13] study and the Schneiders[20] study. This may indicate that these two tests are the hardest to rate consistently due to the array of joints and segments utilized to complete the tasks. Additionally, based on the current utilization of interval scoring using the 0-3 scale, it may be difficult to discern where movement failure occurs, and difficult to describe or quantify such failures with so few scoring choices. Interestingly, Onate et al[17] found the poorest inter-rater reliability on the ASLR with Kappa values for the novices and the trained raters ranging between 0.34-0.44, whereas the raters in the present study demonstrated 100% agreement on this performance test. The authors are unable to determine why this dramatic difference in agreement existed between studies, both of which used raters with a variety of skill and expertise. It may be that experience in test scoring and repetition of scoring procedures may affect the ability to score the FMS reliably. Future studies should account for and report rater qualifications and experience in greater detail to allow for comparisons across studies. This could include certifications, and number of test batteries scored over time.

Limitations

A limitation to this investigation is that each novice was compared to a single expert, but this was consistent with the study design that intended to examine percent agreement between raters of differing training and experience.

Another limitation to the present study was that there were no restrictions placed upon how many times reviewers could view the videotaped records (in two planes of movement) of subjects test performances. Video analysis may have provided the raters in the current study the potential to apply greater scrutiny or critical interpretation to their scoring. The authors acknowledge that video viewing may explain the increased scrutiny of the current raters, and does not mimic the normal utilization of this testing system that typically occurs in real time, with a single viewing of a performance of a test maneuver.

The authors are aware of a proposed 100-point scoring system that eliminates the difficulty encountered with use of an interval based scoring system.[21] Use of such a scoring system may stimulate additional assessments of reliability of the FMS™ scoring, providing for greater discrimination between scores as well as offering detail as to when movement failures occur.

CONCLUSION

In conclusion, the FMS™ is growing in popularity and utilization by fitness and rehabilitation professionals for functional screening of athletes, patients, and clients. Total FMS™ scores appear to be able to be scored reliably between trained raters while the individual tests vary in their ability to be rated reliability. The current results agree are similar to those found by previous authors, which suggest that the composite FMS™ test battery can be used confidently by trained raters to assess fundamental movement patterns and arrive at a total score; whereas the current study suggests the reliability of individual tests scores may be less appropriate for assessing function in isolation, especially when using a heterogenous group of raters.

REFERENCES

1. Cook EG. *Athletic body in Balance: Optimal movement skills and conditioning for performance.* Champaign, IL: Human Kinetics, 2004.
2. Cook EG, Burton L, Hoogenboom BJ. The use of fundamental movements as an assessment of function-Part 1. *N Am J Sports Phys Ther.* 2006;1(2):62-72.
3. Cook EG, Burton L, Hoogenboom BJ. The use of fundamental movements as an assessment of function-Part 2. *N Am J Sports Phys Ther*. 2006;1(3):132-139.
4. Vaughn DW. Isolated knee pain: A case report highlighting regional interdependence. *J Orthop Sports Phys Ther.* 2008;38(10):616-623.
5. Wainner RS, Whitman JM, Cleland JA, Flynn TW. Regional interdependence: A musculoskeletal examination model whose time has come. *J Orthop Sports Phys Ther.* 2007; 37:658-660.
6. Bullock-Saxton JE. Local sensation changes and altered hip muscle function following severe ankle sprain. *Phys Ther.* 1994;74:17-28.
7. Bullock-Saxton JE, Janda, V, Bullock, MI. The influence of ankle sprain injury on muscle activation during hip extension. *Int J Sports Med*. 1994;15:330-334.
8. Cholewicki J, Green HS, Polzhofer GK, Galloway MT, Shah RA, Radebold A. Neuromuscular function in athletes following recovery from a recent acute low back injury. *J Orthop Sports Phys Ther*. 2002;32:568-575.
9. Leetun DR, Ireland ML, Willson JD, Ballantyne BT, Davis IM. Core stability measures as risk factors for lower extremity injury in athletes. *Med Sci Sports Exerc.* 2004;36(6): 926-934.
10. Nadler SF, Malanga GA, Bartoli LA, Feinberg JH, Prybicien M, Deprince M. Hip muscle imbalance and low back pain in athletes: Influence of core strengthening. *Med Sci Sports Exerc.* 2002;34:9-16.
11. Zazulak, BT, Hewett TE, Reeves NP, Goldberg B, Cholewicki J. Deficits in neuromuscular control of the trunk predict knee injury risk. *Am J Sports Med.* 2007;35(7): 1123-1130.
12. Nadler SF, Malanga GA, DePrince M, Stitik TP, Feinberg JH. The relationship between lower extremity injury, low back pain, and hip muscle strength in male and female collegiate athletes. *Clin J Sports Med.* 2000;10(2):89-97.
13. Minick KI, Kiesel KB, Burton L, Taylor A, Plisky P, Butler RJ. Inter-rater reliability of the functional movement screen. *J Strength Cond Res.* 2010;24(2):479-486.
14. Chorba RS, Chorba DJ, Bouillon LE, Overmyer CA, Landis JA. Use of a functional movement screening tool to determine injury risk in female collegiate athletes. *N Am J Sports Phys Ther.* 2010;5(2): 47-54.

15. Okada T, Huxel KC, Nesser TW. Relationship between core stability, functional movement, and performance. *J Strength Cond Res.* 2011;25(1):252-261.
16. Kiesel K, Plisky PJ, Voight ML. Can serious injury in professional football be predicted by a preseason functional movement screen? *N Am J Sports Phys Ther.* 2007;2(3):147-152.
17. Onate JA, Dewey T, Kollock RO, Thomas KS, Van Lunen BL, DeMaio M, Ringleb SI. Real-time intersession and interrater reliability of the functional movement screen. *J Strength Condit Res.* 2013;27(4):978-981.
18. Smith CA, Chimera NJ, Wright NJ, Warren M. Interrater and intrarater reliability of the Functional Movement Screen. *J Strength Condit Res.* 2013;27(4):982-987.
19. Gribble PA, Brigle J, Pietrosimone BG, Pfile KR, Webster KA. Intrarater reliability of the Functional Movement Screen. *J Strength Condit Res.* 2012;26(2):408-415.
20. Schneiders AG, Davidsson A, Horman E, Sullivan SJ. Functional Movement Screen™ normative values in a young, active population. *Int J Sports Phys Ther.* 2011;6(2):75-82.
21. Butler RJ, Plisky PJ, Kiesel KB. Interrater reliability of videotaped performance on the Functional Movement Screen using the 100-point scoring scale. *Ath Train Sports Health Care.* 2012;4(3):103-109.